REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	SEPT 2004	USARIEM TECHNICAL REPORT

4. TITLE AND SUBTITLE
RELIABILITY AND VALIDITY OF DEVICES FOR A LIFE SIGN DETECTION SYSTEM

5. FUNDING NUMBERS

6. AUTHOR(S)
B.A. BEIDLEMAN, W.T. THARION, M.J. BULLER, R.W. HOYT, AND B.J. FREUND

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute of Environmental Medicine
Biophysics and Biomedical Modeling Division
Kansas Street
Natick, MA 01760

8. PERFORMING ORGANIZATION REPORT NUMBER
T05-01

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, MD 21702

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
Life sign detection system (LSDS) devices were developed by the Center for the Integration of Medicine and Innovative Technology (CIMIT, Cambridge, MA, USA), Hidalgo Research Co. (Swavesey, Cambridge, UK), Sarcos Research Co. (Salt Lake City, UT, USA), and VivoMetrics, Inc. (Ventura, CA, USA) for use in the Future Force Warrior (FFW) system. This study evaluated the reliability and validity of these four LSDS devices for measuring heart rate (HR) and respiration rate (RR) against criterion devices (e.g., 3-lead ECG and metabolic cart, respectively). Eight male volunteers (mean ± SD, 21±3 yr, 76±9 kg) completed four days of testing in which HR and RR were collected from the four LSDS devices and criterion devices at the same time for ~4 h while engaged in low, medium, and high intensity activities on two occasions. The change in both HR and RR (i.e., error scores) between the two trials (trial 1-trial 2) and two devices (device-criterion) was calculated and analyzed using a 2-way repeated measures ANOVA. Bland-Altman plots showed the dispersion of individual reliability and validity error scores. All devices demonstrated adequate reliability based on hypothesis testing but Bland-Altman plots revealed tighter prediction intervals for both HR (beats/min) and RR (breaths/min), respectively, for CIMIT (±15.9; ±5.8) and VivoMetrics (±13.0;±9.9) compared to Sarcos (±54.0;±9.4) and Hidalgo (±86.6;±14.2). All devices, except Sarcos, demonstrated adequate validity for both HR and RR based on hypothesis testing but Bland-Altman plots revealed tighter prediction intervals for HR (beats/min) and RR (breaths/min), respectively, for VivoMetrics (±11.6;±21.4) compared to CIMIT (±19.6;±16.5) and Hidalgo (±86.2;±15.7). In conclusion,VivoMetrics was the most reliable and valid LSDS device for measurements of HR and RR. CIMIT was the next most reliable and valid device. The Sarcos and Hidalgo devices should not be evaluated further in their current configuration based on the poor reliability and validity of HR and RR measurements. Even though the VivoMetrics device was the most reliable and valid device, the form factor was not acceptable to the Soldier in the field. Thus, a future LSDS system may need to utilize components of each system to meet all the needs of the Soldier.

14. SUBJECT TERMS
Life sign detection system, ambulatory monitoring, heart rate, respiration rate, validity, reliability

15. NUMBER OF PAGES
40

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
U	U	U	U

USARIEM TECHNICAL REPORT T05-01

RELIABILITY AND VALIDITY OF DEVICES FOR A LIFE SIGN DETECTION SYSTEM

Beth A. Beidleman
William J. Tharion
Mark J. Buller
Reed W. Hoyt
Beau J. Freund

Biophysics and Biomedical Modeling Division

September 2004

U.S. Army Research Institute of Environmental Medicine
Natick, MA 01760-5007

Approved for public release; distribution is unlimited. The opinions or assertions contained herein are the private views of the author(s) and are not be construed as official or as reflecting the views of the Army or the Department of Defense.

Human subjects participated in these studies after giving their free and informed voluntary consent. The investigators have adhered to the policies for protection of human subjects as prescribed in Army Regulation 70-25 and USAMRMC Regulation 70-25 on the use of volunteers in research.

Citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement or approval of the products or services of these organizations.

DTIC AVAILABILITY NOTICE

Qualified requesters may obtain copies of this report from Commander, Defense Technical Information Center (DTIC), Ft Belvoir, Virginia @ 1-800-225-3842

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

BACKGROUND

This research effort supports a milestone under Science and Technology Objective (STO) E (III.ME.2004.01), Warfighter Physiological Status Monitoring-Initial Capability (WPSM-IC). Commanders and medics increasingly require remote, real-time access to basic physiological information about their Soldiers in order to make better decisions regarding the early identification, location, and triage priority of casualties. Historically, it is estimated that early first aid intervention may save up to 20% of all battlefield fatalities (5,6). A life sign detection system (LSDS) could provide combat care providers with essential vital sign information that would assist them in providing rapid first aid to battlefield casualties. Sophisticated reliable and valid output information from the LSDS could lead to reduced morbidity and mortality of both medics and Soldiers by facilitating an appropriate medical response. Although these LSDS devices represent the most sophisticated technology currently available to the U.S. Army, the devices have never been tested for reliability and validity. If the data provided by the LSDS is reliable and valid then more appropriate decisions may be made regarding triage priority. Thus, the military needs to identify a LSDS device that will reliably provide the combat medic with valid life sign information about the Soldier under operational conditions.

ACKNOWLEDGEMENTS

The dedicated and professional efforts of Mr. John Ames, Ms. Laurie Blanchard, SSG Dan Ditzler, Ms. Mary Anne Fawke, Mr. Anthony Karis, Mr. Stephen Mullen, Ms. Jane Simpson, Mr. Leonard Sousa, Ms. Karen Speckman and Dr. Robert Wallace supporting the collection and analysis of the data are acknowledged and greatly appreciated. The dedication and efforts of the test volunteers in completing this study are also acknowledged and appreciated.

EXECUTIVE SUMMARY

Life sign detection system (LSDS) devices were developed by the Center for the Integration of Medicine and Innovative Technology (CIMIT, Cambridge, MA, USA), Hidalgo Research Co. (Swavesey, Cambridge, UK), Sarcos Research Co. (Salt Lake City, UT, USA), and VivoMetrics, Inc. (Ventura, CA, USA) for use in the Future Force Warrior system. This study evaluated the reliability and validity of these four LSDS devices for measuring heart rate (HR) and respiration rate (RR) against criterion devices (e.g., 3-lead ECG and metabolic cart, respectively). Eight male volunteers (mean ± SD, 21±3 yr, 76±9 kg) completed four days of testing in which HR and RR were collected from the four LSDS devices and criterion devices at the same time for ~ 4 h while engaged in low, medium, and high intensity activities on two occasions. The change in both HR and RR (i.e., error scores) between the two trials (trial 1−trial 2) and two devices (device−criterion) was calculated and analyzed using a 2-way repeated measures ANOVA. Bland-Altman plots showed the dispersion of the individual reliability and validity error scores. All devices demonstrated adequate reliability based on hypothesis testing but Bland-Altman plots revealed tighter prediction intervals for both HR (beats•min^{-1}) and RR (breaths•min^{-1}), respectively, for CIMIT (±15.9 and ±5.8) and VivoMetrics (±13.0 and ±9.9) compared to Sarcos (±54.0 and±9.4) and Hidalgo (±86.6 and ±14.2). All devices, except Sarcos, demonstrated adequate validity for both HR and RR based on hypothesis testing but Bland-Altman plots revealed tighter prediction intervals for HR (beats•min^{-1}) and RR (breaths•min^{-1}), respectively, for VivoMetrics (±11.6 and ±21.4) compared to CIMIT (±19.6 and 16.5) and Hidalgo (±86.2 and 15.7). In conclusion, VivoMetrics was the most reliable and valid LSDS device for measurements of HR and RR. CIMIT was the next most reliable and valid device. The Sarcos and Hidalgo devices should not be evaluated further in their current configuration based on the poor reliability and validity of HR and RR measurements. Even though the VivoMetrics device was the most reliable and valid device, the form factor was not acceptable to the Soldier in the field. Thus, a future LSDS system may need to utilize components of each system to meet all the needs of the Soldier.

INTRODUCTION

In order to successfully incorporate any piece of equipment into a military system, it is essential to determine whether the piece of equipment works appropriately by providing reliable and valid data under various activity levels. Reliability refers to the reproducibility of data collected from equipment in repeated tests on the same individuals (1,13). Validity refers to the agreement between the value of a measurement and its true value (25,5). The concepts of reliability and validity are related. For example, measurements can be reliable but not valid, but a valid measurement must be reliable. Several different methods for assessing reliability and validity exist (1). The five most common measures of reliability are the change in mean between tests, standard error of measurement (SEM), coefficient of variation (CV), which is the SEM expressed as a percent of the subject's mean score, Bland-Altman 95% limits of agreement, and intraclass correlation coefficient (ICC). The five most common measures of validity are the change in mean between devices, the estimation (i.e., regression) equation, the standard error of the estimate (SEE), Bland-Altman 95% limits of agreement, and the Pearson correlation coefficient. Since arguments for and against each method exist depending on the expert quoted, multiple kinds of evidence should be presented in order to best interpret reliability and validity data (9,13).

Reliability and validity studies are routinely done on new pieces of equipment that claim to provide physiological data that is comparable to accepted criterion methods of physiological data collection (3,16-18,20,24). Given that airway, breathing, and circulation are the basic ABC's of maintaining life as published by the American Heart Association (10), heart rate (HR) and respiration rate (RR) are critical physiological measures of life sign status in the Soldier. Although a 3-lead electrocardiograph (ECG) for measuring heart rate (HR) and a metabolic cart for measuring respiration rate (RR) have been previously validated (15,21-23) these criterion methods are not practical for operational use by a Soldier in the field because of their size, weight, and inability to be moved. Therefore, ambulatory "wear and forget" life sign detection system (LSDS) devices were developed as a collaborative effort between the U.S. Army Research Institute of Environmental Medicine (USARIEM) and the following four companies: (a) Center for the Integration of Medicine and Innovative Technology (CIMIT, Cambridge, MA, USA), (b) Hidalgo Research Co. (Swavesey, Cambridge, UK), (c) Sarcos Research Co. (Salt Lake City, UT, USA) and (d) VivoMetrics, Inc. (Ventura, CA, USA). The LSDS devices provided by these four companies purport to provide HR and RR data equivalent to previously-validated

criterion methods. However, none of these LSDS devices have been tested for reliability and validity on human subjects under a variety of physical activity levels. An independent lab validation study, therefore, was conducted by USARIEM in order to more clearly identify the most acceptable LSDS device for possible insertion into the Future Force Warrior (FFW) military system. Furthermore, since engaging in military activities is likely to affect the signal quality of recorded physiological data (11), quantifying the percent of time that unusable signals are generated in various levels of activity (i.e., low, medium, and high) will enable manufacturers to improve and enhance current technologies in order to better meet the Soldier's needs.

OBJECTIVES

The first objective of this study was to determine the reliability and validity of HR and RR collected from four LSDS devices under a wide range of low (i.e., lying, sitting, standing), medium (i.e., sit-ups, push-ups, jumping jacks) and high (i.e., walking, running) intensity activities that are part of normal military duties. Reliability was determined via test-retest methods while validity was determined by comparing the HR and RR data collected from the LSDS devices to the criterion device. The second objective of this study was to determine whether any of the LSDS devices provided acceptable data under low, medium, and high activity levels. Acceptable data was defined as a HR and RR within (a) ± 3% of the criterion measure 90% of the time during low activity, (b) ±4% of the criterion measure 80% of the time during medium activity, and (c) ±5% of the criterion measure 70% of the time during high activity.

METHODS

SUBJECTS

Eight male soldiers with a mean (±SD) age, body weight, and height of 21±3 yr, 76±9 kg, 175±5 cm, respectively, enrolled in this study. Each gave written and verbal acknowledgment of their informed consent and was made aware of their right to withdraw without prejudice at any time. Investigators adhered to the policies for protection of human subjects as prescribed in Army Regulation 70-25, and the research was conducted in adherence with the provisions of 45 CFR Part 46.

PROTOCOL

Design

Eight Soldiers wore one of the four LSDS devices and criterion devices at the same time for ~ 4 h each day while engaging in low, medium, and high intensity activities on two occasions (Table 1).

Table 1. Schedule of Testing Activities

Activity Intensity	Activity	Time of Activity	Total Time
Low	1. Lie on cot 2. Sit on cot 3. Stand on ground 4. Break 5. Repeat	1. ~10 min 2. ~10 min 3. ~10 min 4. ~10 min 5. ~40 min	~ 80 min
Medium	1. Walk on treadmill at 3.0 mph 2. Sit-ups, push-ups, jumping jacks 3. Break 4. Repeat	1. ~15 min 2. ~10 min 3. ~10 min 4. ~35 min	~ 70 min
High	1. Run 1.0 mi on treadmill at 6.0 mph 2. Break 3. Repeat	1. ~10 min 2. ~10 min 3. ~20 min	~ 40 min

The HR and RR were collected every 15-30 s over the ~4 h of testing from both the LSDS and criterion device. Each day of testing was followed by at least one day of non-testing and no exercise such that volunteers were never tested more than three times in one week. Volunteers were tested at the same time of the day during the four days of testing. The study design was counterbalanced so that each Soldier wore a different LSDS device on each day of testing to control for order effects (Table 2). Soldiers were randomly assigned a testing order.

Table 2. Counterbalanced Experimental Design

Volunteer /Device	VivoMetrics	Hidalgo	Sarcos	CIMIT
1	Day 1	Day 4	Day 2	Day 3
2	Day 2	Day 3	Day 4	Day 1
3	Day 3	Day 2	Day 1	Day 4
4	Day 4	Day 1	Day 3	Day 2
5	Day 1	Day 4	Day 2	Day 3
6	Day 2	Day 3	Day 4	Day 1
7	Day 3	Day 2	Day 1	Day 4
8	Day 4	Day 1	Day 3	Day 2

Description of LSDS Devices

Vivometrics LifeShirt™ is an FDA-approved noninvasive ambulatory recording device that continuously acquires and stores respiration, ECG, and body position data on a data card stored within a portable battery powered electronic recorder worn on the body. The LifeShirt system consists of a LifeShirt garment, data cable, ECG electrodes, recorder battery, data cards, electronic recorder, battery charger and recorder case.

Hidalgo Thoracic Sensor is a prototype chest-mounted sensor, worn centrally on the thoracic area and comprised of two parts. The first is a wireless re-usable sensor electronics module that measures ECG-based heart rate and respiration rate. The second part is an adhesive electrode disposable laminate, comprising upper and lower sections, both of which have a removable adhesive. The two parts are joined by metal studs, which allow the upper and lower parts to move freely with respect to one another.

Sarcos Integrated Sensor Unit is a prototype physiologic data acquisition system for non-invasive monitoring of vital signs. The Integrated Sensor Unit measures ECG-based heart rate, presence/absence of breathing, breathing rate, body motion and position, and axillary temperature. The sensor is configured in a form of a chest belt and attached to the chest using an elastic strap.

CIMIT Soft Chest Strap is a prototype device intended for use in collecting physiologic status in real time. The components of the CIMIT Soft Chest Strap include (a) conductive rubber ECG contacts encased in a polymeric compound, (b) sensor case made of a rigid polyurethane material which is designed for commercial use with a

number of products requiring contact with human skin, and (c) the polyester and Velcro chest strap, which contains no latex.

Description of Criterion Devices

The Schiller Cardiovit AT-6 ECG machine (Schiller Inc., Baar, Switzerland) was utilized as the criterion device for measuring HR, and has met ECG instrument specifications of the American Heart Association (2). The SensorMedics 2900 automated metabolic cart system (SensorMedics Corporation, Yorba Linda, CA) was utilized as the criterion device for measuring RR, and has been previously shown to be a reliable and valid instrument for measuring respiration (15,22).

Environmental Conditions

All testing was performed in the Doriot Climatic chamber located in Natick, MA, at the Natick Soldier Systems Center. The temperature and relative humidity in the 60x10x14 ft climatic chamber was maintained at 21±2 °C and 45±5 %, respectively for all testing. The environmental conditions did not differ on any of the four testing days.

STATISTICS

The change in both HR and RR (i.e., error scores) between the two trials (trial1−trial2) and devices (device−criterion) was calculated. Given the known effects of previous activity on elevating HR and RR, a negative value was expected for the change in mean between trials during medium and high activity. However, the negative bias should be similar for all devices given that the activity and time between activities was standardized on each day of testing. A two-way repeated measures ANOVA (device x activity) was utilized to determine the reliability and validity of each LSDS device during each activity by comparing the reliability and validity error scores across devices and activities. Individual error scores of zero indicated that the reliability (trial1−trial 2) and validity (device−criterion) of the device was high. Bland-Altman plots were constructed to show the dispersion of the individual reliability error scores (trial1−trial2) and validity error scores (device−criterion) (7). In this manner, the mean error score was illustrated, and the 95% limits of agreement (prediction interval) were depicted. Individual error scores that have a tight prediction interval around zero signified a more reliable and valid device. In addition, the coefficient of variation (CV) and intraclass correlation coefficient (ICC) were calculated for both HR and RR

measured on each LSDS device and criterion method during each of the eight activities as further measurements of reliability. The standard error of the estimate (SEE) and Pearson correlation coefficient were also calculated for both HR and RR measured on each LSDS device during each of the eight activities as further measurements of validity. The SEE is a measure of the accuracy of predictions made with a regression line utilizing the LSDS device measurement of HR or RR (i.e., independent variable) to predict the criterion measurement of HR or RR (i.e., dependent variable). In other words, any time you use a regression line to estimate a subject's true heart rate from the device, there will be an error. This error, expressed as a standard deviation, is the SEE. Just like the CV, defining an acceptable SEE depends on how far the estimated measurement can vary from the true measurement without affecting decisions. For instance, in this study, a SEE for HR of 4 beats•min^{-1} would suggest that the predicted HR would vary by plus or minus 2 x (SEE) or 8 beats•min^{-1} 95% of the time. Given the acceptability criteria that HR should be within ± 5% of the gold standard and the maximum HR during running was ~160 beats•min^{-1}, a SEE for HR of 4 beats•min^{-1} would be reasonable. Following the same reasoning, a SEE for RR of 1 breath•min^{-1} would be reasonable. The percentage of time a usable signal (i.e., HR and RR) was detected was determined by dividing the total number of received signals by the total possible number of signals. The percent deviation of HR and RR from the LSDS device compared to the criterion device was calculated as device minus criterion divided by the criterion. This measurement was used for defining acceptability. All data are presented as mean ± SD. Statistical significance was set at $p<0.05$.

RESULTS

RELIABILITY

Figure 1 demonstrates the overall change in HR mean between trials (i.e. overall error score) for each device combining all eight activities. None of the error scores were significantly different from zero. Bland-Altman plots demonstrated that two devices (CIMIT and VivoMetrics) provided more reliable measurements of HR than the others (Figure 2). For convenience, the y-axes are standardized to highlight the differences in reliability between devices. Table 3 demonstrates that the overall prediction intervals (beats•min^{-1}) for both CIMIT (±15.9) and VivoMetrics (±13.0) were similar to that for the criterion measure (±15.1) but were ~four times greater for Sarcos (±54.0) and ~six times greater for Hidalgo (±86.6). In addition, the overall CV (%) for CIMIT (5.1) and VivoMetrics (4.2) was comparable to the criterion measure (5.4) but much greater for

Sarcos (17.5) and Hidalgo (27.8) (Table 3). When individual activity CVs (%) were examined, CIMIT was comparable to the criterion measure during all activities, Hidalgo and Sarcos were comparable to the criterion measure only during low activities, and VivoMetrics was comparable to the criterion measure during all activities except sit-ups and push-ups. The ICC reached significance (i.e., high reliability) in six of eight activities for CIMIT, three of eight activities for Hidalgo, three of eight activities for Sarcos, and three of eight activities for VivoMetrics (Table 3).

Figure 3 demonstrates the overall change in RR mean between trials (i.e., overall error score) for each LSDS device combining all eight activities. None of the error scores were significantly different from zero. Bland-Altman plots demonstrated that two devices (CIMIT and VivoMetrics) provided more reliable measurements of RR than the others (Figure 4). Table 4 demonstrates that the overall prediction interval (breaths•min^{-1}) for CIMIT (±5.8) and VivoMetrics (±6.4) was similar to that for the criterion measure (±5.2) but was ~two times greater for Sarcos (±9.4) and Hidalgo (±14.2). In addition, the overall CV (%) for CIMIT (8.3) and VivoMetrics (9.9) was comparable to criterion (6.5) but much greater for Sarcos (26.5) and Hidalgo (41.8) (Table 4). When individual activity CVs were examined (Table 4), CIMIT was comparable to the criterion measure during medium and high intensity activities but not low intensity activities, Hidalgo was comparable to the criterion measure only during lying, Sarcos was not comparable to the criterion measure in any activity and VivoMetrics was comparable to the criterion measure during all activities except sit-ups. The ICC reached significance in seven of eight activities for CIMIT, four of eight activities for Hidalgo, four of eight activities for Sarcos, and six of eight activities for VivoMetrics.

Although cut-off reliability scores were not defined prior to the conduct of this protocol, a device can not be valid if it does not demonstrate adequate reliability. Thus, if a device failed both CV and ICC expectations, their reliability could be considered questionable. The devices that failed these expectations for HR and RR and thus demonstrated questionable reliability are highlighted in light gray in Tables 3 and 4, respectively.

VALIDITY

Figure 5 demonstrates the device minus criterion HR mean between trials (i.e., overall error score) for each device combining all eight activities. The validity error score for Sarcos was significantly different from zero indicating that it significantly

underestimated the HR compared to the criterion device. The remaining devices did not significantly over- or under-estimate HR. Bland-Altman plots demonstrated that CIMIT and VivoMetrics provided more valid measurements of HR than Hidalgo (Figure 6). The overall prediction interval (beats•min^{-1}) for Hidalgo (±86.2) was ~ four to seven times greater than the overall prediction interval for CIMIT (±19.6) and VivoMetrics (±11.6), respectively. When individual activity SEEs were examined (Table 5), CIMIT performed poorly during sit-ups, push-ups, and jumping jacks, Hidalgo performed poorly during all activities, Sarcos performed poorly during medium and high intensity activities, and VivoMetrics performed well during all activities except push-ups. The Pearson correlation coefficients were significant for six of eight activities for CIMIT, zero of eight activities for Hidalgo, four of eight activities for Sarcos, and eight of eight activities for VivoMetrics.

Figure 7 illustrates the device minus criterion RR mean between trials (i.e., overall error score) for each device combining all eight activities. The validity error score for Sarcos was significantly different from zero indicating that it significantly underestimated the RR compared to the criterion. The remaining devices did not significantly over- or under-estimate RR. Bland-Altman plots demonstrated similar prediction intervals (breaths•min^{-1}) for CIMIT (±16.5), VivoMetrics (±21.4), and Hidalgo (±15.7) (Figure 8). When individual activity SEEs were examined (Table 6), CIMIT, Hidalgo, and Sarcos performed poorly during medium and high intensity activities while VivoMetrics only performed poorly during sit-ups, push-ups, and jumping jacks. The Pearson correlation coefficient for RR was significant for four of eight activities for CIMIT, zero of eight activities for Hidalgo, zero of eight activities for Sarcos, and five of eight activities for VivoMetrics.

Although SEE and Pearson correlation coefficient expectations were not defined prior to the conduct of the protocol, if a device failed to meet fair and reasonable expectations for these measures, then their validity could be considered questionable. The devices that failed these expectations for HR and RR and thus demonstrated questionable validity are highlighted in light gray in Tables 5 and 6, respectively.

ACCEPTABILITY

Acceptable data was defined as HR and RR within (a) ± 3% of the criterion measure 90% of the time during low intensity activity, (b) ±4% of the criterion measure 80% of the time during medium intensity activity, and (c) ±5% of the criterion measure 70% of the time during high intensity activity. The three medium intensity activities (i.e., sit-ups, push-ups, and jumping jacks) were combined into one category for the

percentage of time a usable signal was detected due to the short time of each of these activities (~1-2 min). Figure 9 demonstrates that all devices met the percentage of time criteria except during lying and sitting for Hidalgo and during standing for CIMIT. Although the percentage of time acceptability criteria were generally met, Figure 9 demonstrates that Hidalgo had the lowest percentage of time a usable signal was detected while VivoMetrics had the highest. CIMIT and Sarcos were in the middle as far as percentage of time a usable signal was received, with CIMIT performing slightly better on HR and Sarcos performing better on RR. Figure 10 shows that VivoMetrics was the only device that met acceptability criteria for HR during all eight activities. CIMIT met the acceptability criteria for HR during all activities except push-ups. Sarcos met the acceptability criteria for HR during all activities except sit-ups and push-ups but there was a large percent deviation from the criterion device during these activities. Hidalgo met the acceptability criteria for HR only during three activities: sit-ups, jumping jacks, and running. Figure 10 also demonstrates that VivoMetrics met the acceptability criteria for RR in six out of eight activities. CIMIT and Hidalgo met the RR acceptability criteria in only one of eight activities (i.e., walking). Sarcos failed to meet the acceptability criteria for RR during any of the activities. Devices that failed to meet the acceptability criteria are highlighted in dark gray in Tables 5 and 6.

Table 3. Reliability of Heart Rate (HR) Measured During Two Consecutive Trials by Four Devices (i.e., CIMIT, Hidalgo, Sarcos, and VivoMetrics) and a Criterion Device During Eight Activities.

Device	Activity	Change in Mean Between Trials (beats•min^{-1})	Coefficient of Variation (%)	Limits of Agreement (beats•min^{-1})	Intraclass Correlation Coefficient
CIMIT	Lying	4.7±2.7	3.0	±5.4	0.89*
	Sitting	4.0±3.3	3.5	±6.6	0.89*
	Standing	5.2±4.9	4.7	±9.8	0.72*
	Sit-ups	-7.2±8.5	5.4	±17.0	0.58*
	Push-ups	-16.3±12.4	7.8	±24.8	0.20
	Jumping Jacks	-5.6±8.6	3.9	±17.2	0.83*
	Walking	-18.6±5.6	3.9	±11.2	-0.27
	Running	2.1±2.7	1.2	±5.4	0.93*
Hidalgo	Lying	0.1±6.3	6.1	±12.6	0.97*
	Sitting	-0.1±5.4	5.3	±10.8	0.97*
	Standing	3.2±6.9	5.7	±13.8	0.97*
	Sit-ups	-10.5±13.3	8.3	±26.6	0.55
	Push-ups	-15.8±19.8	10.6	±39.6	0.51
	Jumping Jacks	-10.2±29.5	13.6	±59.0	0.26
	Walking	31.3±39.2	20.6	±78.4	0.19
	Running	-5.4±40.2	17.7	±80.4	0.22
Sarcos	Lying	2.7±5.0	5.6	±10.0	0.76*
	Sitting	1.3±7.1	8.0	±14.2	0.60*
	Standing	4.4±4.8	4.6	±9.6	0.46
	Sit-ups	7.9±29.0	23.7	±58.0	0.31
	Push-ups	-1.8±33.6	27.7	±67.2	-0.31
	Jumping Jacks	-14.0±19.2	9.2	±38.4	-0.38
	Walking	-8.9±24.3	16.6	±48.6	-0.47
	Running	9.6±14.2	6.2	±28.4	0.62*
VivoMetrics	Lying	4.9±6.0	6.9	±12.0	0.39
	Sitting	2.6±3.6	3.9	±7.2	0.84*
	Standing	5.6±3.0	3.0	±6.0	0.73*
	Sit-ups	-11.1±6.7	4.2	±13.4	0.55
	Push-ups	-15.0±21.2	10.0	±42.4	0.42
	Jumping Jacks	-5.6±7.3	3.4	±14.6	0.85*
	Walking	-19.7±6.0	4.2	±12.0	-0.24
	Running	4.5±16.9	7.4	±33.8	0.05
Criterion	Lying	1.8±5.2	5.8	±10.4	0.70*
	Sitting	1.7±3.8	4.2	±7.6	0.78*
	Standing	3.8±5.7	5.4	±11.4	0.45
	Sit-ups	-9.6±2.9	1.9	±5.8	0.95*
	Push-ups	-7.1±9.6	4.8	±19.2	0.90*
	Jumping Jacks	-5.4±10.9	5.1	±21.8	0.69*
	Walking	-6.9±11.0	7.3	±22.0	0.03
	Running	4.8±4.7	2.1	±9.4	0.94*

*P<0.05, white indicates acceptable reliability; light gray indicates questionable reliability

Table 4. Reliability of Respiration Rate (RR) Measured During Two Consecutive Trials by Four Devices (i.e., CIMIT, Hidalgo, Sarcos, and VivoMetrics) and a Criterion Device During Eight Activities.

Device	Activity	Change in Mean Between Trials (breaths•min^{-1})	Coefficient of Variation (%)	Limits of Agreement (breaths•min^{-1})	Intraclass Correlation Coefficient
CIMIT	Lying	-1.2±2.3	14.7	±4.6	0.93*
	Sitting	-0.5±1.5	7.3	±3.0	0.75*
	Standing	0.3±2.9	15.7	±5.8	0.86*
	Sit-ups	-2.1±4.9	13.3	±9.8	0.54
	Push-ups	0.1±2.9	7.5	±5.8	0.73*
	Jumping Jacks	-1.5±5.5	9.5	±11.0	0.70*
	Walking	-4.3±1.8	4.9	±3.6	0.64*
	Running	0.5±2.9	5.1	±5.8	0.86*
Hidalgo	Lying	-0.8±1.4	8.8	±2.8	0.97*
	Sitting	-0.7±2.3	13.7	±4.6	0.90*
	Standing	-0.3±2.0	9.9	±4.0	0.98*
	Sit-ups	-3.4±3.8	12.8	±7.6	0.43
	Push-ups	-2.6±4.1	13.3	±8.2	0.77*
	Jumping Jacks	-4.1±5.9	18.3	±11.8	-0.02
	Walking	4.6±7.5	21.1	±15.0	0.36
	Running	-1.5±9.8	30.3	±19.6	-0.31
Sarcos	Lying	-1.4±2.4	85.7	±4.8	0.38
	Sitting	0.2±1.6	117.0	±3.2	0.38
	Standing	-0.3±1.3	55.6	±2.6	0.79*
	Sit-ups	-0.4±2.5	13.9	±5.0	0.75*
	Push-ups	-0.5±3.7	18.3	±7.4	0.83*
	Jumping Jacks	-0.5±5.9	36.2	±11.8	0.50
	Walking	-2.0±10.7	93.3	±21.4	0.20
	Running	3.7±7.1	39.8	±14.2	0.65*
VivoMetrics	Lying	-0.0±2.4	10.2	±4.8	0.47
	Sitting	-0.2±1.2	5.1	±2.4	0.80*
	Standing	1.1±2.3	9.3	±4.6	0.74*
	Sit-ups	0.9±8.6	17.6	±17.2	0.67*
	Push-ups	-2.7±4.5	8.4	±9.0	0.87*
	Jumping Jacks	-1.8±2.1	2.9	±4.2	0.98*
	Walking	-4.9±1.3	3.5	±2.6	-0.27
	Running	-0.9±5.1	8.5	±10.2	0.81*
Criterion	Lying	-1.3±1.3	5.6	±2.6	0.78*
	Sitting	-0.6±1.2	5.6	±2.4	0.74*
	Standing	-1.2±2.3	9.1	±4.6	0.79*
	Sit-ups	-1.4±1.3	5.5	±2.6	0.93*
	Push-ups	-0.7±2.6	6.0	±5.2	0.70*
	Jumping Jacks	-0.0±2.6	4.8	±5.2	0.93*
	Walking	-3.6±2.6	7.2	±5.2	0.81*
	Running	-1.3±3.9	6.2	±7.8	0.92*

*P<0.05, white indicates acceptable reliability; light gray indicates questionable reliability

Table 5. Validity of Heart Rate (HR) Measured During Two Consecutive Trials by Four Devices (i.e., CIMIT, Hidalgo, Sarcos, and VivoMetrics) During Eight Activities.

Device	Activity	Device - Criterion (beats•min^{-1})	Standard Error of the Estimate	Limits of Agreement (beats•min^{-1})	Pearson Correlation Coefficient
CIMIT	Lying	-0.6± 0.3	0.3	±0.6	0.99*
	Sitting	-0.1± 1.1	1.2	±2.2	0.99*
	Standing	-0.3± 0.8	0.8	±1.6	0.99*
	Sit-ups	-1.4± 5.1	5.3	±10.2	0.88*
	Push-ups	-19.5± 19.6	20.4	±39.2	0.36
	Jumping Jacks	8.1± 17.9	16.8	±35.8	0.41
	Walking	-0.1± 0.5	0.5	±1.0	0.99*
	Running	-0.9± 2.4	2.6	±4.8	0.96*
Hidalgo	Lying	6.9± 20.2	8.2	±40.4	0.62
	Sitting	6.1± 16.1	8.6	±32.2	0.66
	Standing	11.1± 32.1	8.3	±64.2	-0.03
	Sit-ups	-1.5± 14.1	12.6	±28.2	0.52
	Push-ups	-16.9± 23.0	11.7	±46.0	0.09
	Jumping Jacks	2.9± 18.2	16.5	±36.4	0.52
	Walking	30.8± 34.2	10.6	±68.4	-0.33
	Running	-0.2± 20.0	10.9	±40.0	0.53
Sarcos	Lying	-1.5± 1.3	0.8	±2.6	0.99*
	Sitting	-1.9± 2.3	1.7	±4.6	0.95*
	Standing	-0.9± 2.3	2.3	±4.6	0.89*
	Sit-ups	-21.3± 20.8	9.6	±41.6	0.08
	Push-ups	-57.0± 16.0	17.2	±32.0	0.64
	Jumping Jacks	-3.8± 12.5	13.0	±25.0	0.34
	Walking	-2.3± 15.4	11.9	±30.8	-0.29
	Running	-0.4± 5.7	4.0	±11.4	0.96*
VivoMetrics	Lying	-3.1± 1.3	1.3	±2.6	0.97*
	Sitting	-2.5± 0.8	0.9	±1.6	0.99*
	Standing	-2.7± 1.2	1.1	±2.4	0.99*
	Sit-ups	1.6± 3.4	3.4	±6.8	0.96*
	Push-ups	-0.2± 13.3	12.9	±25.8	0.76*
	Jumping Jacks	0.1± 2.4	2.4	±4.8	0.99*
	Walking	-1.3± 0.7	0.8	±1.4	0.99*
	Running	-1.9± 3.4	3.0	±6.8	0.97*

*P<0.05, white indicates acceptable validity; light gray indicates questionable validity, dark gray indicates unacceptable validity

Table 6. Validity of Respiration Rate (RR) Measured During Two Consecutive Trials by Four Devices (i.e., CIMIT, Hidalgo, Sarcos, and VivoMetrics) During Eight Activities.

Device	Activity	Device - Criterion (breaths•min^{-1})	Standard Error of the Estimate	Limits of Agreement (breaths•min^{-1})	Pearson Correlation Coefficient
CIMIT	Lying	-4.9 ± 7.1	1.9	±14.2	-0.14
	Sitting	-1.0 ± 1.0	0.9	±2.0	0.85*
	Standing	-5.1 ± 7.4	2.8	±14.8	-0.63
	Sit-ups	-6.2 ± 8.2	7.1	±16.4	-0.05
	Push-ups	-2.9 ± 1.7	1.6	±3.4	0.88*
	Jumping Jacks	2.5 ± 5.0	5.2	±10.0	0.71*
	Walking	0.1 ± 3.9	3.4	±7.8	0.62
	Running	-4.4 ± 6.3	6.3	±12.6	0.79*
Hidalgo	Lying	-6.1 ± 7.0	1.6	±14.0	-0.41
	Sitting	-4.3 ± 5.7	1.2	±11.4	-0.47
	Standing	-2.5 ± 8.8	1.5	±17.6	0.03
	Sit-ups	-9.4 ± 7.6	5.3	±15.2	-0.37
	Push-ups	-8.4 ± 4.7	3.9	±9.4	0.69
	Jumping Jacks	-13.9 ± 8.7	6.8	±17.4	-0.38
	Walking	-0.9 ± 7.2	3.6	±14.4	-0.08
	Running	-19.5 ± 7.2	7.1	±14.4	0.04
Sarcos	Lying	-14.7 ± 2.4	2.2	±4.8	0.33
	Sitting	-14.9 ± 1.6	1.6	±3.2	0.37
	Standing	-16.3 ± 2.7	2.4	±5.4	0.08
	Sit-ups	-15.7 ± 8.6	8.2	±17.2	-0.11
	Push-ups	-14.2 ± 9.7	8.1	±19.4	-0.05
	Jumping Jacks	-23.4 ± 11.7	8.5	±23.4	-0.43
	Walking	-17.7 ± 7.6	3.1	±15.2	-0.25
	Running	-30.9 ± 8.7	8.1	±17.4	0.45
VivoMetrics	Lying	-0.1 ± 0.7	0.8	±1.4	0.92*
	Sitting	0.2 ± 0.8	0.8	±1.6	0.91*
	Standing	0.1 ± 1.3	1.3	±2.6	0.92*
	Sit-ups	-0.5 ± 10.0	6.1	±20.0	0.14
	Push-ups	4.1 ± 9.1	5.8	±18.2	0.39
	Jumping Jacks	12.9 ± 13.7	7.7	±27.4	0.30
	Walking	0.8 ± 0.9	1.0	±1.8	0.86*
	Running	0.4 ± 0.3	0.3	±0.6	0.99*

*P<0.05, white indicates acceptable validity; light gray indicates questionable validity, dark gray indicates unacceptable validity

Figure 1. Heart Rate Reliability Error Scores Combining Eight Activities

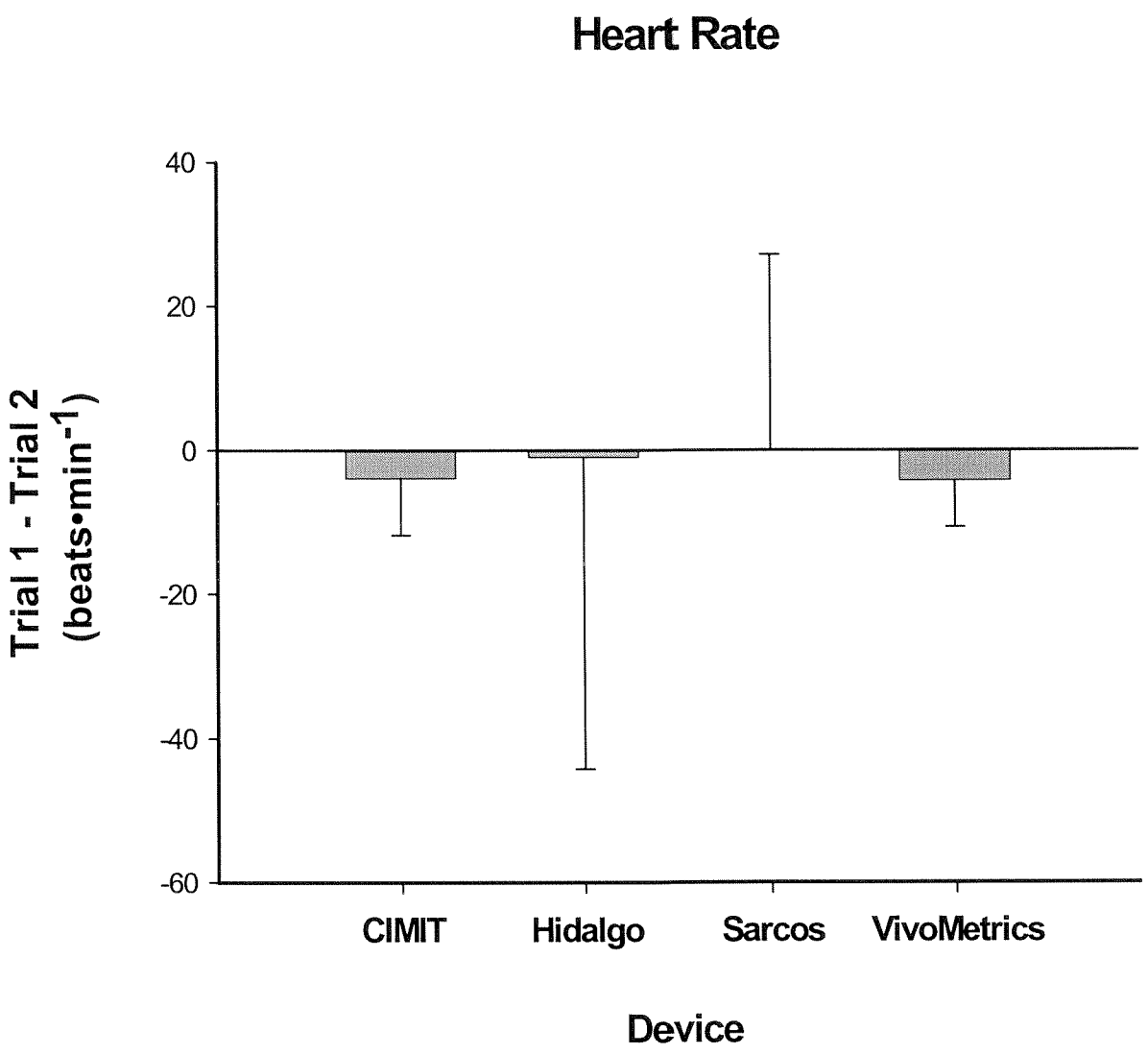


Figure 2. Bland-Altman Plots for Individual Heart Rate Reliability Error Scores Combining Eight Activities

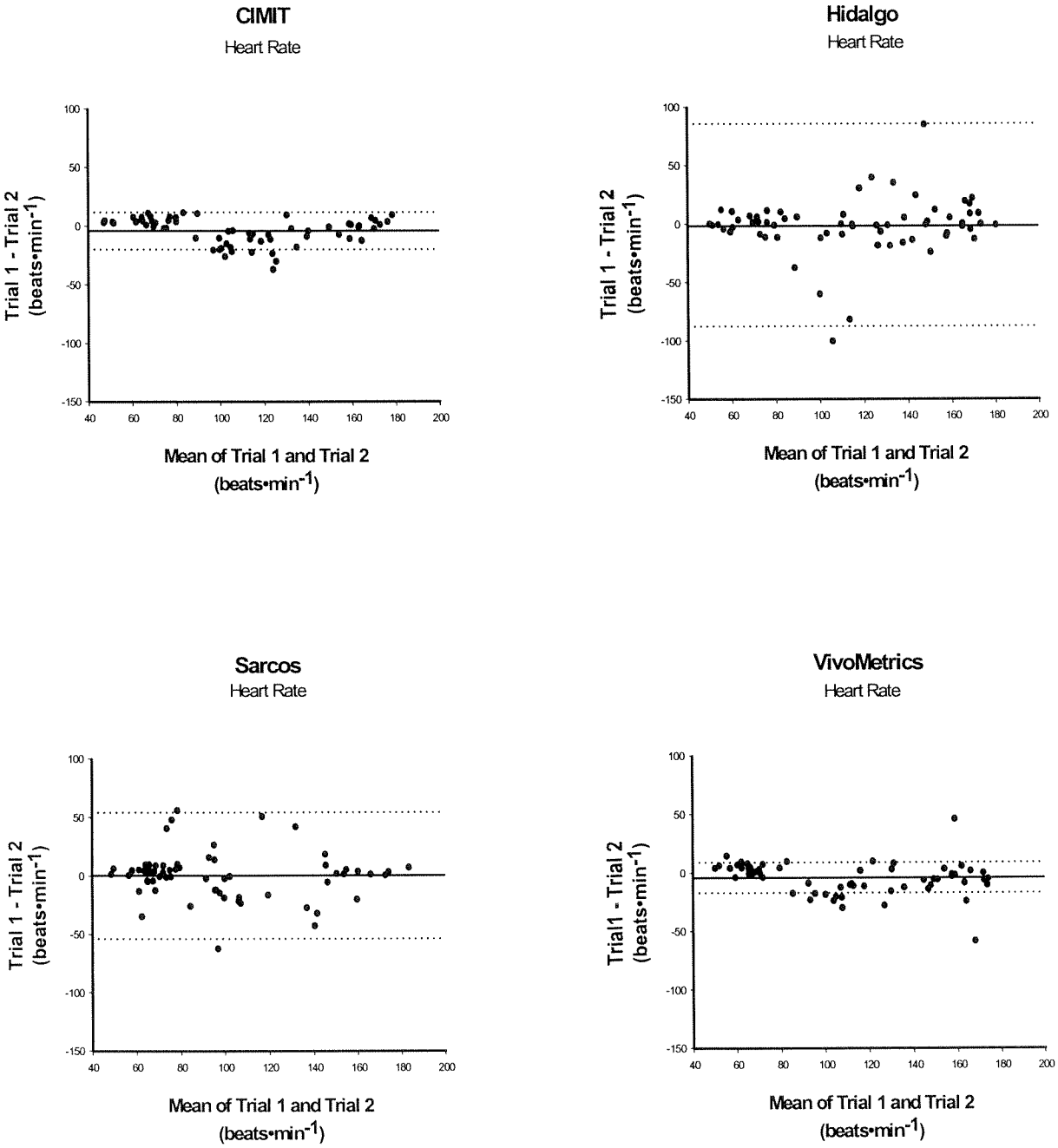


Figure 3. Respiration Rate Reliability Error Scores Combining Eight Activities

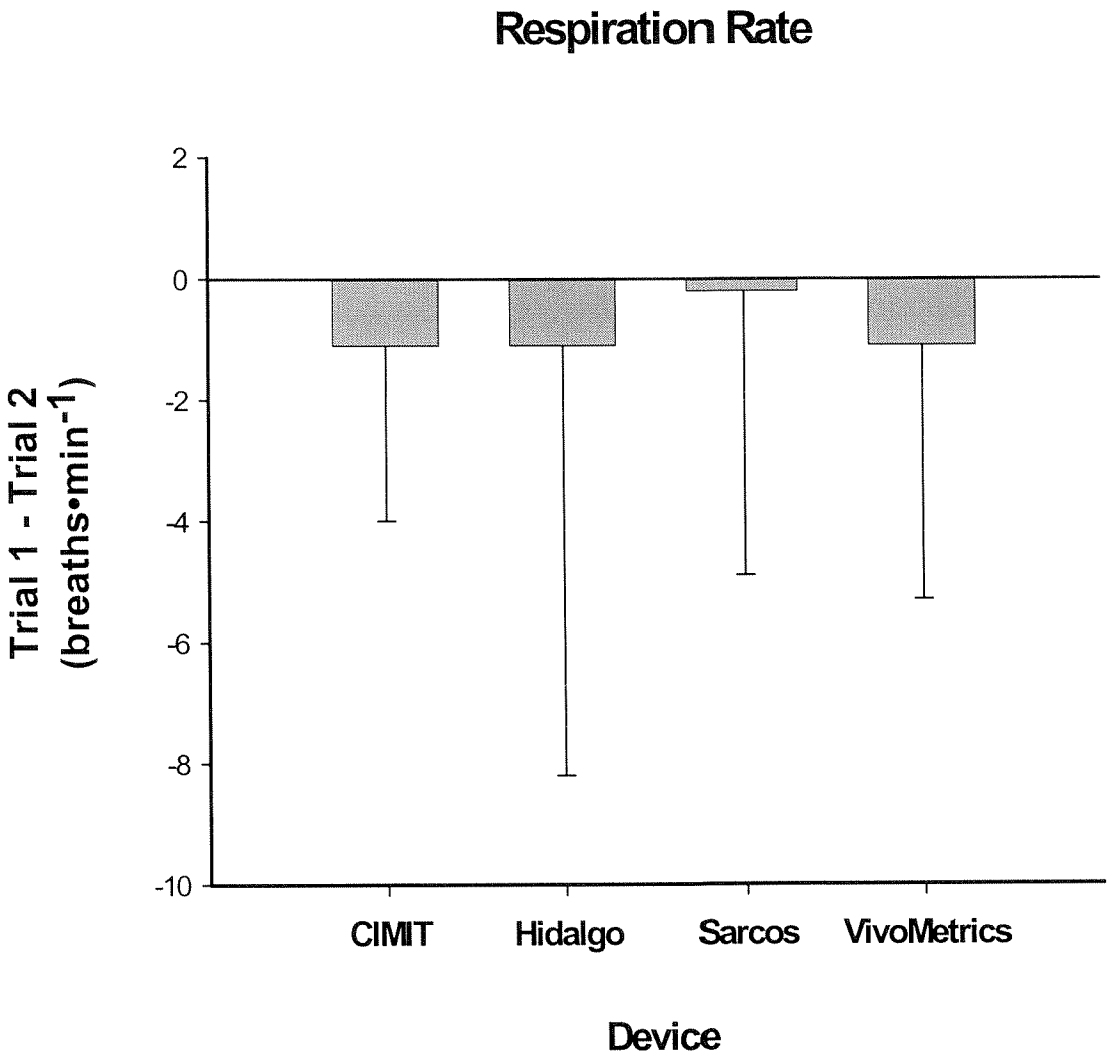


Figure 4. Bland-Altman Plots for Individual Respiration Rate Reliability Error Scores Combining Eight Activities

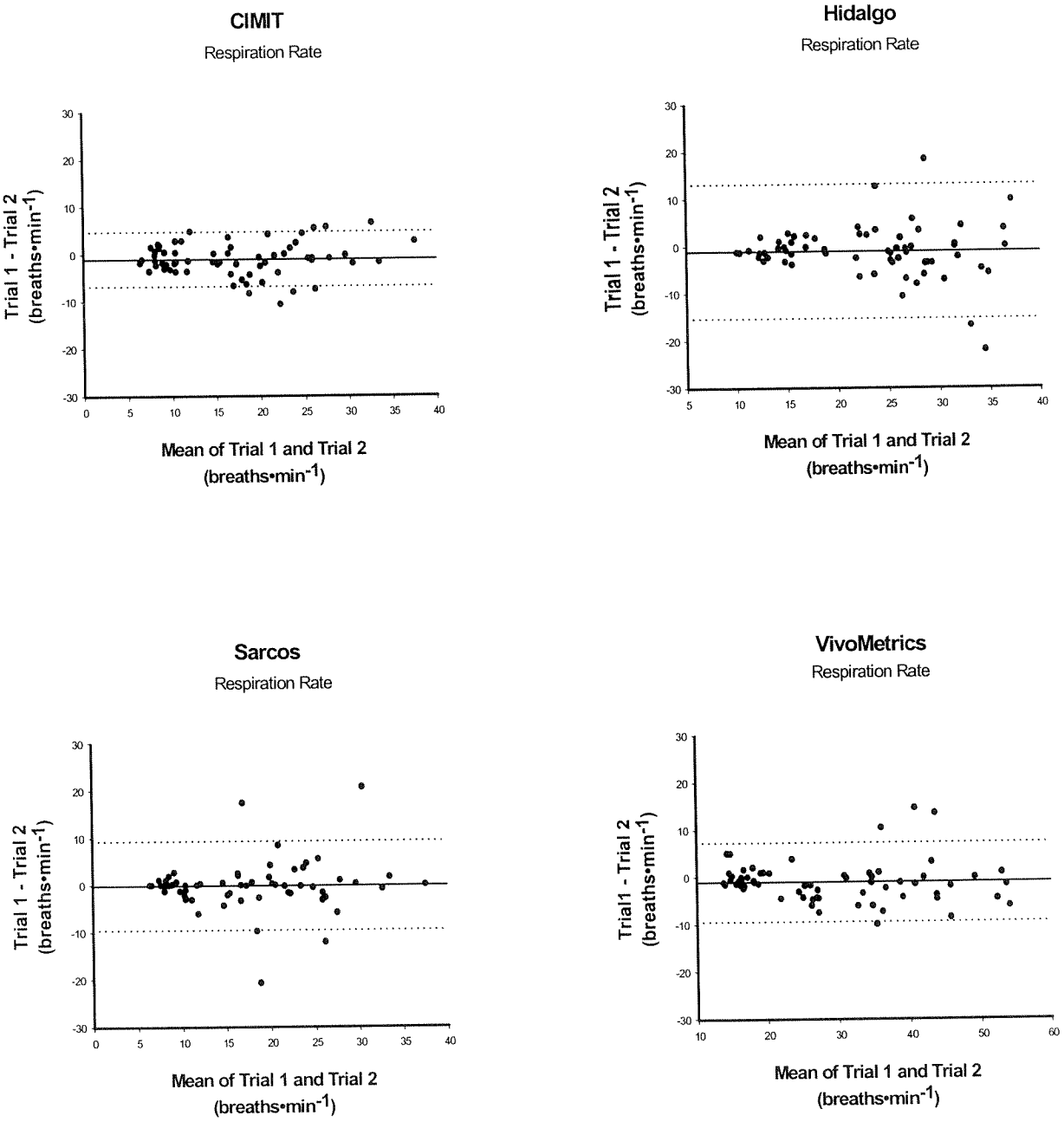
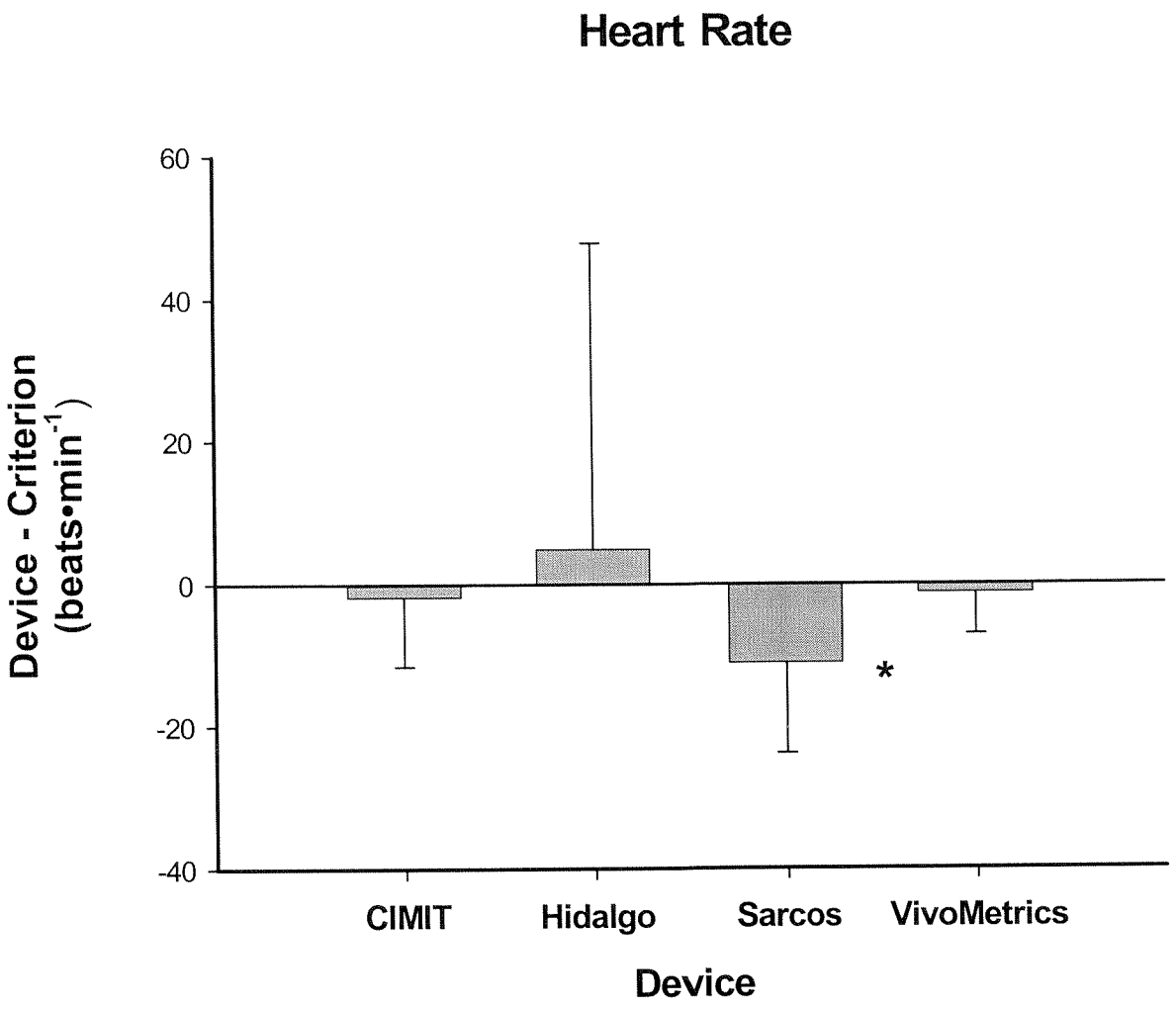


Figure 5. Heart Rate Validity Error Scores Combining Eight Activities



* P<0.05

Figure 6. Bland-Altman Plots for Individual Heart Rate Validity Error Scores Combining Eight Activities

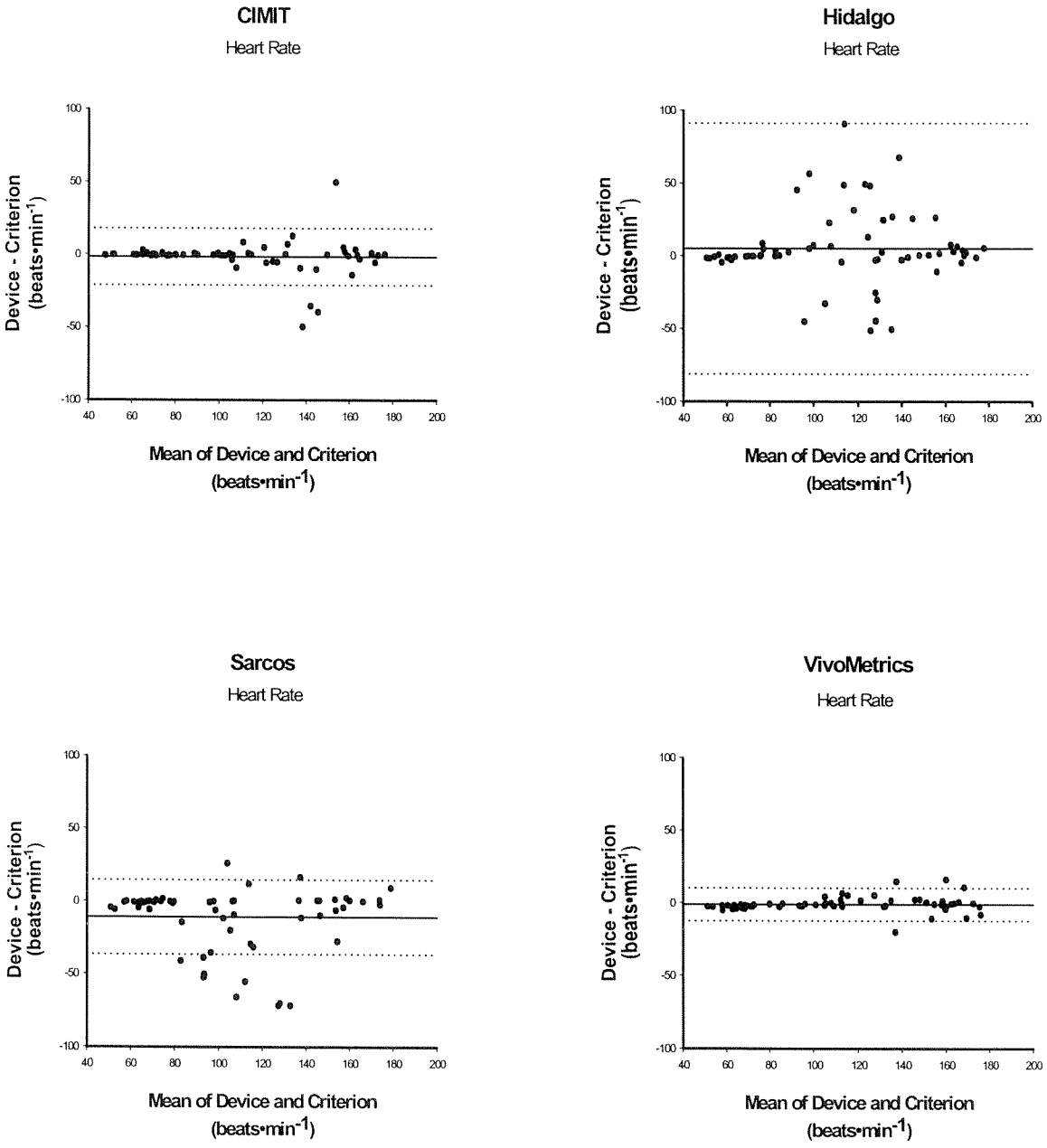
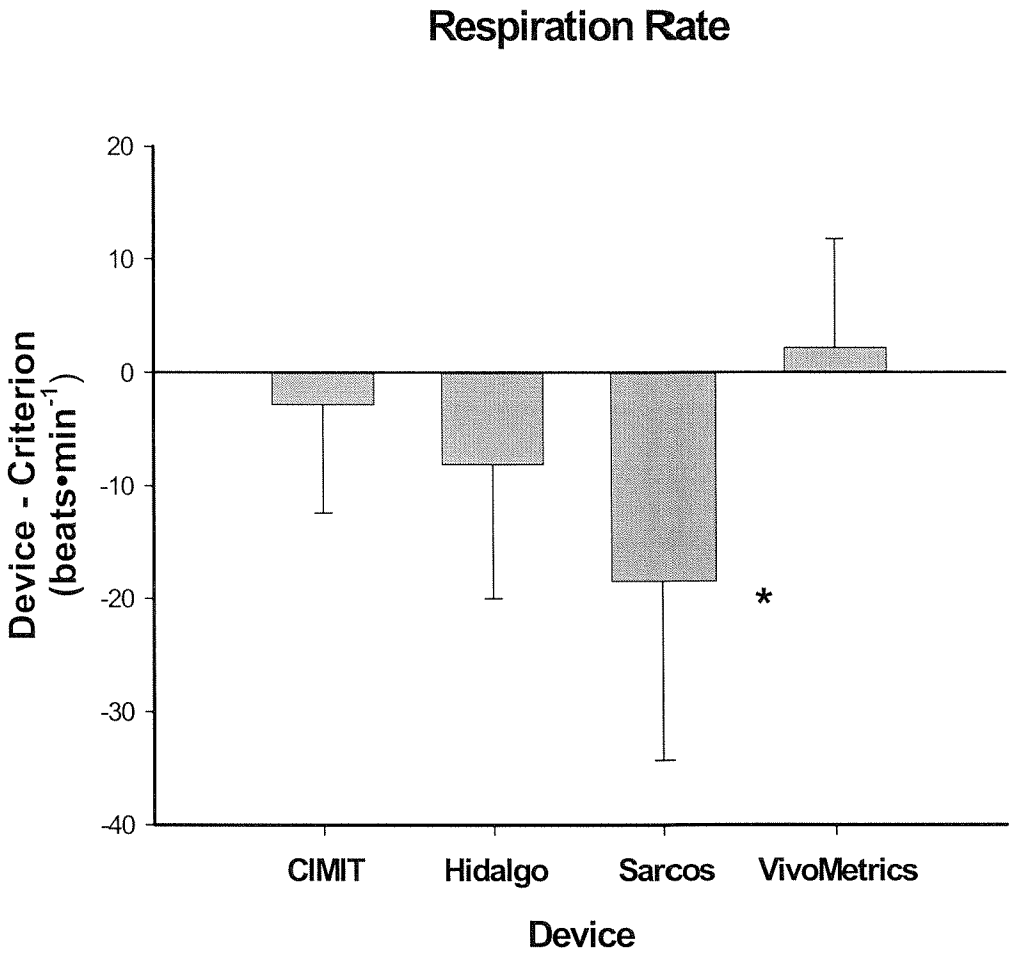


Figure 7. Respiration Rate Validity Error Scores Combining Eight Activities



* P<0.05

Figure 8. Bland-Altman Plots for Individual Validity Respiration Rate Error Scores Combining Eight Activities

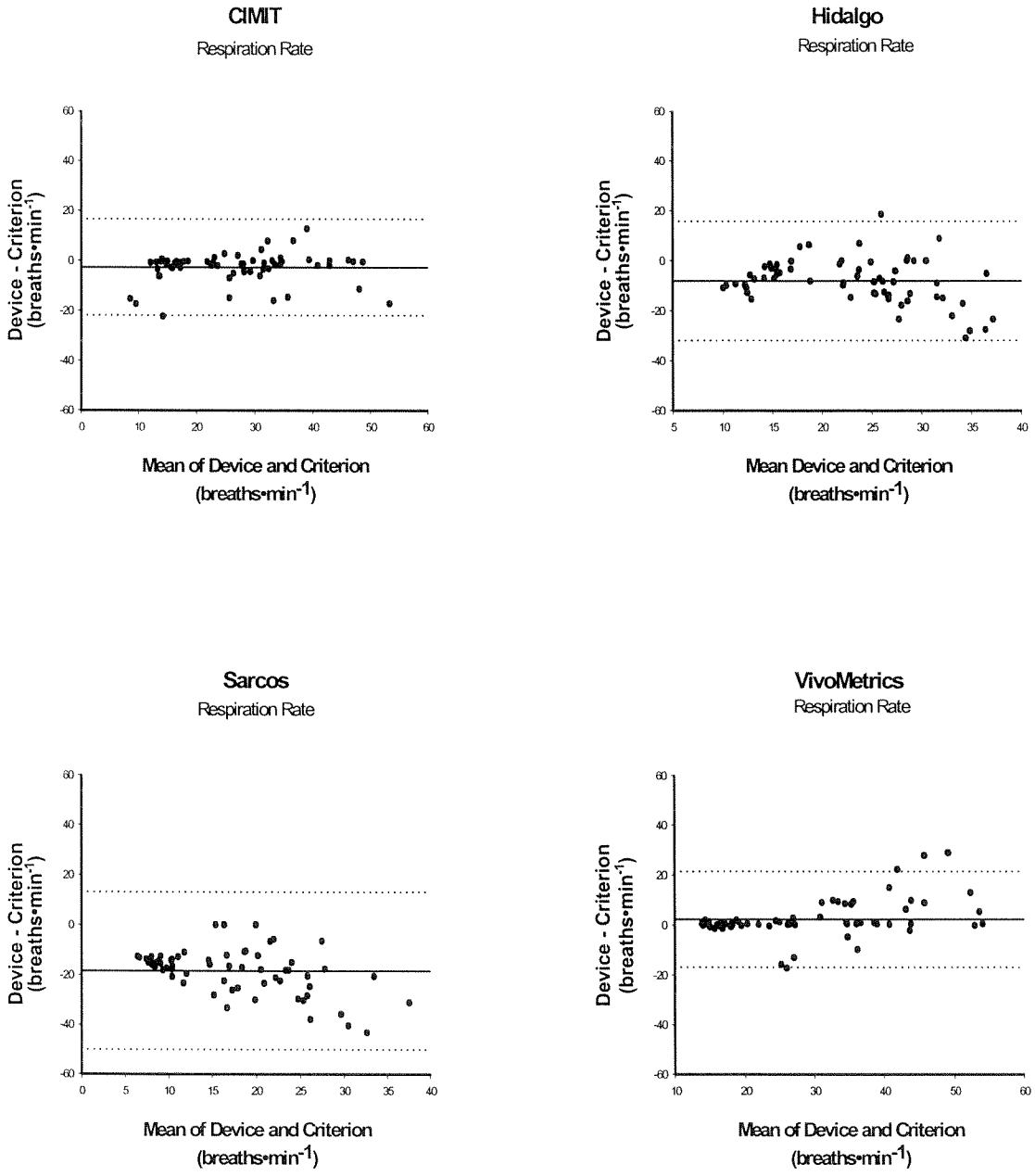
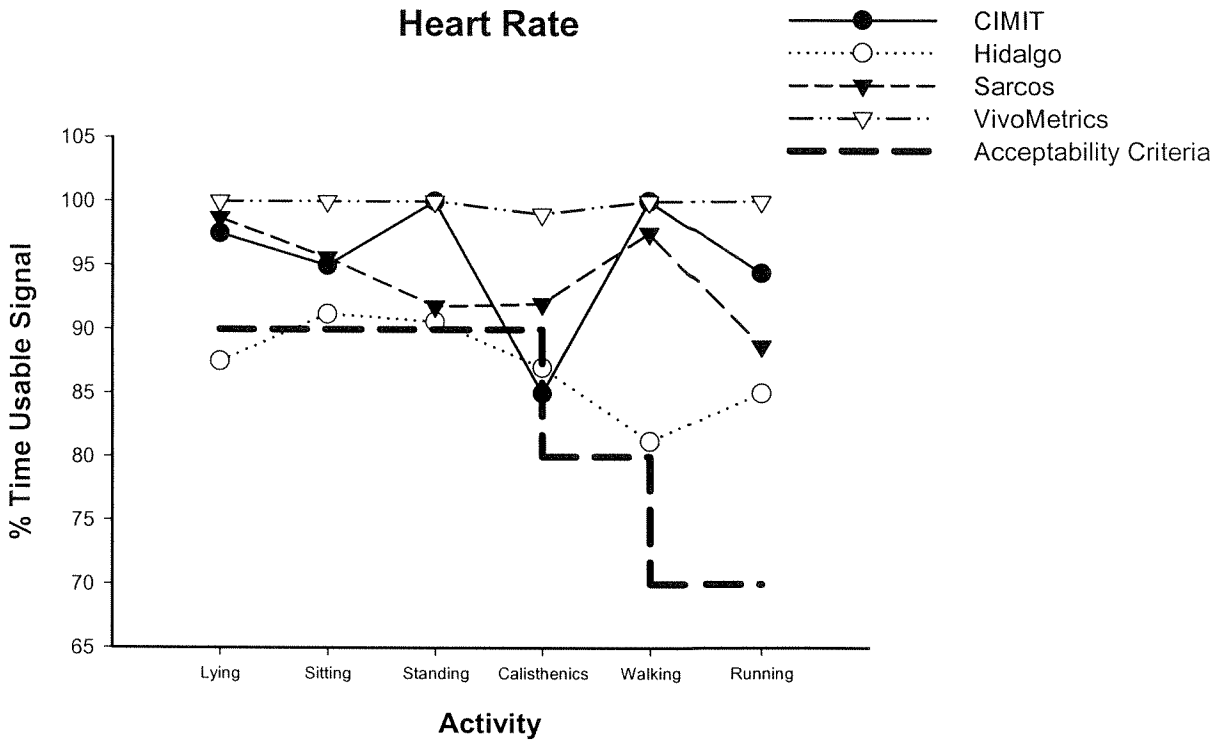


Figure 9. Percentage of Time Usable Signal Detected



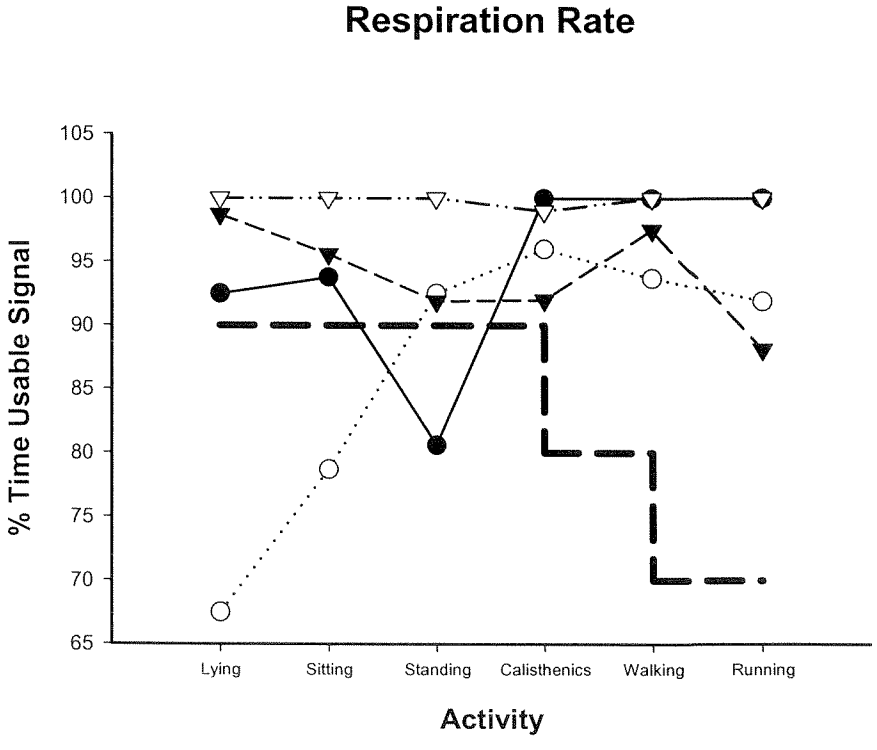
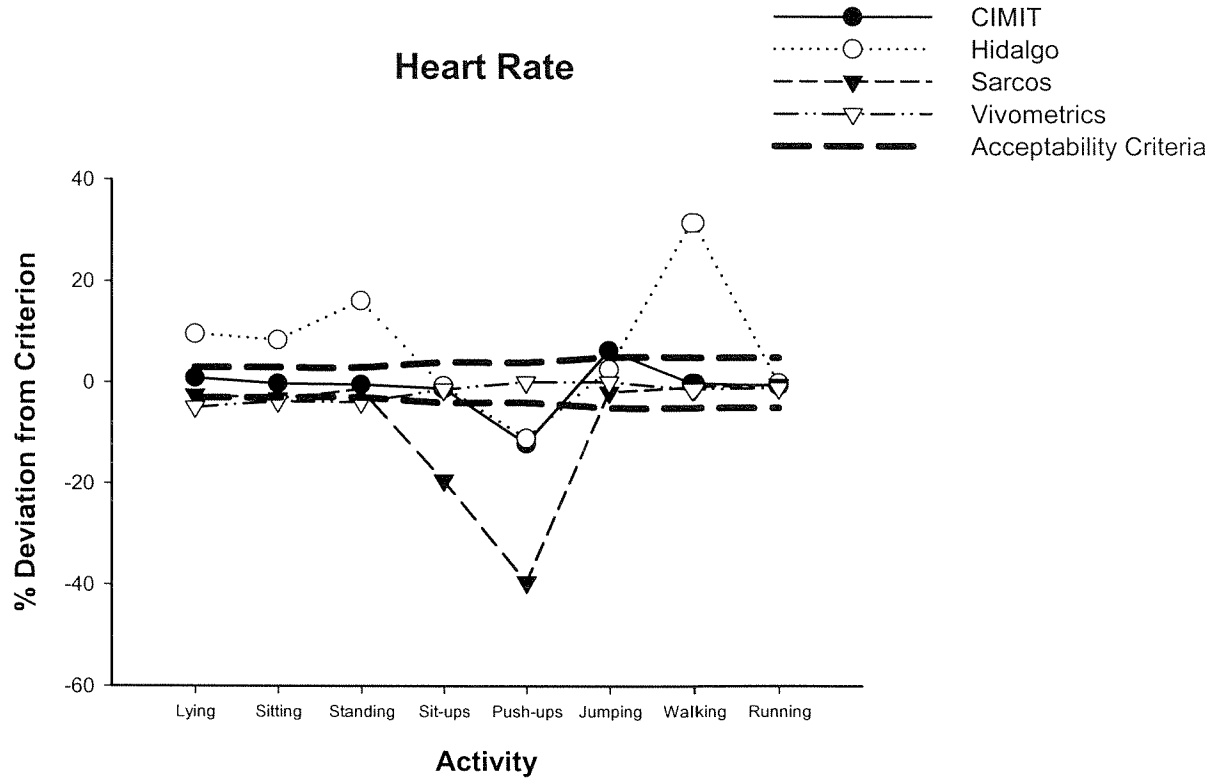
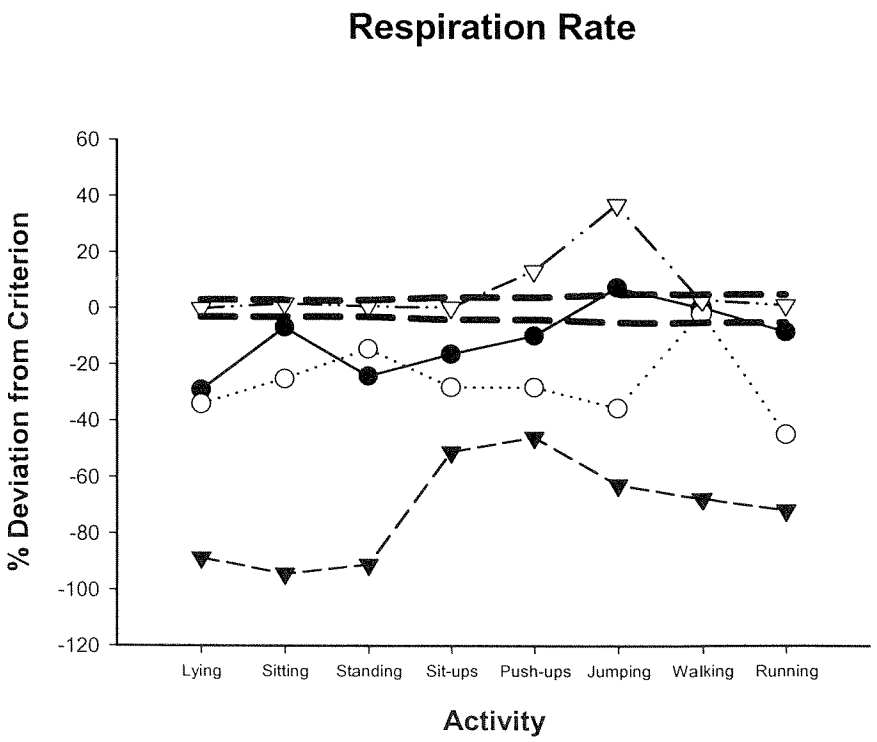


Figure 10. Percent Deviation from Criterion Device





DISCUSSION

The purpose of this study was to evaluate the reliability and validity of four LSDS devices (CIMIT, Hidalgo, Sarcos, and VivoMetrics) for measuring heart rate (HR) and respiration rate (RR) against criterion devices (i.e., 3-lead ECG and metabolic cart, respectively). This was done by comparing the HR and RR obtained by these four LSDS devices against the criterion devices during eight activities (i.e., lying, sitting, standing, sit-ups, push-ups, jumping jacks, walking, and running) on two occasions.

The results demonstrated that all of the devices demonstrated adequate reliability by hypothesis testing since none of the error scores for HR and RR between trials (Trial 1 – Trial 2) was significantly different from zero. However, the difficulty with hypothesis testing is that with large standard deviations and low subject numbers, the result is biased to detect no differences between methods (1).

Recent studies measuring reliability and validity of biomedical equipment have employed Bland-Altman 95% limits of agreement (LOA) plots in order to more closely examine individual error scores for biomedical devices (7,16,17,20). Devices that demonstrate individual repeatability error scores that have a tight 95% prediction interval around zero are more reliable devices (4,14). Utilizing Bland-Altman criteria, Figure 2 clearly demonstrates that CIMIT and VivoMetrics were more reliable for HR measurements than Hidalgo and Sarcos given their tight 95% prediction interval around zero. Repeated measurements of HR by Hidalgo and Sarcos could vary by as much as ±87 beats•min^{-1} and ±54 beats•min^{-1}, respectively, of the mean difference in HR between trials. This prediction interval is too large for practical purposes. The CIMIT, VivoMetrics, and criterion devices all had similar prediction intervals of ±15 beats•min^{-1} of the mean difference between trials. Figure 4 also demonstrates that CIMIT, VivoMetrics, and Sarcos were more reliable for RR measurements than the Hidalgo device for similar reasons.

The CV and the ICC have also been recommended as useful tools for analyzing repeatability of biomedical instrumentation (13). The CV, which represents the variability between repeated measurements, has been previously reported for ECG machines (19,23), and is similar to the 5.4% reported for the criterion device in this study. Given the prototype development stage of these devices as a LSDS, accepting a CV that is double (i.e., 10%) the standard CV for measurement of HR would be reasonable. In this scenario, only CIMIT and VivoMetrics met this expectation during all eight activities. Hidalgo did not meet this CV expectation for HR measurements during push-ups, jumping jacks, walking, or running, while Sarcos did not meet this

expectation during sit-ups, push-ups, or walking (Table 3). The CV for measuring RR by the metabolic cart has also been previously reported and is similar to the 6.5% reported in this study (14,15). Again, given the development stage of these devices, doubling the CV (i.e., ~13.0%) for measuring RR between trials would be reasonable. In this scenario, the CV expectation was met by VivoMetrics in seven of eight activities, by CIMIT in four of eight activities, by Hidalgo in three of eight activities, and by Sarcos in zero of eight activities (Table 4).

The ICC has also been widely used as a measure of repeatability between trials versus the Pearson correlation coefficient because systematic errors in measurement are reflected by a reduction in the value of the coefficient (8,12). Utilizing this technique for assessing reliability, none of the devices achieved a significant ICC during all eight activities for either HR or RR. CIMIT demonstrated the greatest number of significant correlation coefficients for both HR and RR.

Considering all four methods of measuring reliability, it is clear that both CIMIT and VivoMetrics were more reliable than Sarcos and Hidalgo. Based on the weight of HR reliability evidence, the devices would demonstrate the following rank order: (a) CIMIT, (b) VivoMetrics, (c) Sarcos, and (d) Hidalgo. Based on the weight of RR reliability evidence, the devices would demonstrate the following rank order: (a) VivoMetrics, (b) CIMIT, (c) Hidalgo, and (d) Sarcos.

Hypothesis testing of the mean difference in HR and RR measured by the LSDS device and criterion device demonstrated that Sarcos was the only device in which HR and RR validity error scores were significantly different from zero. All of the other devices demonstrated acceptable validity error scores.

However, Bland-Altman 95% LOA plots for individual validity HR error scores demonstrate that VivoMetrics provided more valid measures of HR than CIMIT, Hidalgo, or Sarcos. The prediction intervals for CIMIT and Sarcos were double that for Vivometrics while the prediction interval for Hidalgo was approximately eight times wider than the prediction interval for VivoMetrics. The Bland-Altman plots for individual RR validity error scores demonstrate that CIMIT, Hidalgo, and Vivometrics performed similarly while the prediction interval for Sarcos was 30% wider than for the others.

Given that expectations for the SEE and Pearson correlation coefficient were not defined prior to the conduct of the protocol, a device should not be considered unacceptable based on these criteria. Furthermore, the Pearson correlation coefficient does not necessarily measure agreement between two device but rather association between them (12). Nonetheless, if a device failed to meet both of these criteria, then the validity of that device should be considered questionable.

The results in total of the different methods of assessing validity also clearly demonstrate that CIMIT and VivoMetrics are more valid devices for measuring HR and RR than Hidalgo and Sarcos. Based on the weight of HR validity evidence, the devices would demonstrate the following rank order: (a) VivoMetrics, (b) CIMIT, (c) Sarcos, and (d) Hidalgo. Based on the weight of RR validity evidence, the devices would demonstrate the following rank order: (a) VivoMetrics, (b) CIMIT, (c) Hidalgo, and (d) Sarcos.

The acceptability criteria that were defined prior to the conduct of the protocol stated that devices would demonstrate a HR and RR within (a) ± 3% of the criterion measure 90% of the time during low intensity activity, (b) ±4% of the criterion measure 80% of the time during medium intensity activity, and (c) ±5% of the criterion measure 70% of the time during high intensity activity. Although most devices met the percentage of time a usable signal was received criteria during most activities except for Hidalgo during lying and sitting and CIMIT during standing, Figure 9 clearly shows that the Hidalgo device demonstrated decreased performance in this area. The reason for this is that the Hidalgo system was an adhesive-based system that frequently fell off subjects during high intensity activity and failed to adhere properly during low intensity activity. The reasons for the lower percentage of time a usable signal was received for HR from the Sarcos system was that the belt frequently fell down from the expected chest position during high activity. Reasons for the decreased percentage of time a signal was received for RR in the CIMIT system is not entirely known but may have been due to signal transduction from the belt to the receiver. Reasons for the high percentage of time a signal was received for both HR and RR for the VivoMetrics system was most likely due to the constraining vest type design. Although this user design may be appropriate for collecting reliable and valid data, it was not an acceptable user design for the Soldier in the field (4).

Figure 10 shows that VivoMetrics was the only device that met acceptability criteria for HR during all eight activities. CIMIT met acceptability criteria in 6 of 8 activities, Hidalgo in 3 of 8 activities, and Sarcos in 6 of 8 activities. None of the devices met the acceptability criteria for RR during all activities. However, VivoMetrics met the acceptability criteria for RR during five of the eight activities whereas CIMIT and Hidalgo met the acceptability criteria in only one of eight activities. Sarcos failed to meet the acceptability criteria for RR during any of the activities. The results of this study are not surprising given that VivoMetrics is an FDA-certified device. Thus, certain reliability and validity criteria were already met before FDA-certification.

CONCLUSIONS

Based on the reliability and validity results of this study, VivoMetrics was the most reliable and valid device based on several statistical measures. CIMIT was the next most reliable and valid device and provided more valid measurements of HR than RR. The Sarcos and Hidalgo devices should not be evaluated further in their current configuration based on the poor reliability and validity of HR and RR measurements in this study. Even though the VivoMetrics device was the most reliable and valid device, the form factor was not acceptable to the Soldier in the field. Thus, a future LSDS system may need to utilize components of each system to meet all the needs of the Soldier. Future research should also examine other physiological variables related to dead/alive status such as body position, movement, blood pressure, and body temperature to more accurately determine life sign status.

REFERENCES

1. Atkinson, G., and A. M. Nevill. Statistical methods for assessing measurement error (reliability) in variables relevant to Sports Medicine. *Sports Med.* 26: 217-238, 1998.

2. Bailey, J. J., A. S. Berson, Jr. A. Garson, L. G. Horan, P. W. McFarlane, Mortara.D.W., and C. Zywietz. Recommendations for standardization and specifications in automated electrocardigraphy: bandwidth and digital signal processing. A report for health professional by an ad hoc writing group of the Committee on Electrocardiography and Cardiac Electrophysiology of the Council on Clinical Cardiology, American Heart Association. *Circulation* 81: 730-739, 1990.

3. Balmer, J., R. C. R. Davison, D. A. Coleman, and S. R. Bird. The validity of power output recorded during exercise performance tests using a Kingcycle air-braked cycle ergometer when compared with an SRM powermeter. *Int. J. Sports Med.* 21: 195-199, 2000.

4. Beidleman, B. A., Hoyt, R. W., Pearce, F. J., Ditzler, D. T., Ames, J., Speckman, K. L., Blanchard, L. A., Garcia, A., Gaffney, W. P, and Freund, B. J. *User Acceptability of Design Concepts for a Life Sign Detection System.* Natick, MA: USARIEM Technical Report T03-12, 2003.

5. Bellamy, R. F. The causes of death in conventional land warfare: Implications for combat casulaty care research. *Mil. Med.* 149: 55-62, 1984.

6. Bellamy, R. F. Death on the battlefield and the role of first aid. *Mil. Med.* 152: 634-635, 1987.

7. Bland, J. M., and D. G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1: 307-310, 1986.

8. Darras, D. J. Statistical methodology: II. Reliability and validity assessmet in study

design, part A. *Acad. Emerg. Med* 4: 64-71, 1997.

9. Doherty, M., J. Balmer, R. C. R. Davison, L. Robinson, and P. M. Smith. Reliability of a combined 3-min constant load and performance cycling test. *Int. J. Sports Med.* 24: 366-371, 2003.

10. Emergency Cardia Care Committee and Subcommittees, A. H. A. Guidelines for cardiopulmonary resuscitation and emergency cardiac care. *JAMA* 268: 2171-2295, 1992.

11. Hamilton, P. S., M. Curley, and Aimi.R. Effect of adaptive motion-artifact reduction on QRS detection. *Biomed. Instrum. Technol.* 34: 197-202, 2000.

12. Hebert, J. R., and D. R. Miller. The inappropriateness of conventional use of the correlation coefficient in assessing validity and reliability of dietary assessment methods. *Eur. J. Epidemiol.* 7: 339-343, 1991.

13. Hopkins, W. G. Measures of reliability in sports medicine and science. *Sports Med.* 30: 1-15, 2000.

14. Jensen, K., S. Jorgensen, and L. Johansen. A metabolic cart for measurement of oxygen uptake during human exercise using inspiratory flow rate. *Eur. J. Appl. Physiol.* 87: 206, 2002.

15. Macfarlane, D. J. Automated metabolic gas analysis systems. A review. *Sports Med.* 31: 841-861, 2001.

16. McLaughlin, J. E., G. A. King, E. T. Howley, D. R. Bassett Jr., and B. E. Ainsworth. Validation of the COSMED K4 b2 portable metabolic system. *Int. J. Sports Med.* 22: 280-284, 2001.

17. Millet, G. P., C. Tronche, N. Fuster, D. J. Bentley, and R. Candau. Validity and reliability of the Polar S710 mobile cycling powermeter. *Int. J. Sports Med.* 24: 156-

161, 2003.

18. O'Brien, C., R. W. Hoyt, M. J. Buller, J. W. Castellani, and A. J. Young. Telemetry pill measurement of core temperature in humans during active heating and cooling. *Med. Sci. Sports Exerc.* 30: 468-472, 1998.

19. Revill, R. M., and M. D. L. Morgan. Biological quality control for exercise testing. *Thorax* 55: 63-66, 2000.

20. Schneider, P. L., S. E. Crouter, O. Lukajic, and D. R. Bassett Jr. Accuracy and reliability of 10 pedometers for measuring steps over a 400-m walk. *Med. Sci. Sports Exerc.* 35: 1779-1784, 2003.

21. Skinner, J. S., K. M. Wilmore, A. Jaskolska, A. Jaskolski, E. W. Daw, T. Rice, J. Gagnon, A. S. Leon, J. H. Wilmore, D. C. Rao, and C. Bouchard. Reproducibility of maximal exercise test data in the HERITAGE Family Study. *Med. Sci. Sports Exerc.* . 31: 1623-1628, 1999.

22. Unnithan, V. B., J. Wilson, D. Buchanan, J. A. Timmons, and J. Y. Paton. Validation of the sensormedics (S2900Z) metabolic cart for pediatric exercise testing. *Can. J. Appl. Physiol.* 19: 472-479, 1994.

23. Wergel-Kolmert, U., a. Wisen, and B. Wohlfart. Repeatability of measurements of oxygen consumption, heart rate and Borg's scale in men during ergometer cycling. *Clin. Physiol. Funct. Imaging* 22: 261-265, 2002.

24. White, W. B., P. Lund-Johansen, and P. Omvik. Assessment of four ambulatory blood pressure monitors and measurements by clinicians versus intraarterial blood pressure at rest and during exercise. *Am. J. Cardiol.* 65: 60-66, 1990.

25. Zar, J. H. *Biostatistical analyses.* London: Prentice Hall, 1996.